UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF OHIO

JANE ROE, individually and on behalf of all others similarly situated, Plaintiff, vs. INTELLICORP RECORDS, INC., an Ohio corporation, and DOES 1-50, inclusive, Defendant.	Case No.: 1:12-CV-02288-JG Judge: James S. Gwin **REPLY DECLARATION OF HENK VALK IN SUPPORT OF PLAINTIFF'S MOTION FOR CLASS CERTIFICATION**

I, HENK VALK, declare:

1. I am an individual over the age of eighteen residing in Napa County, California. I have been retained by the attorneys for the Plaintiff in this action as a consultant and expert witness on information technology (IT) and database issues, and I am submitting this Declaration on Plaintiff's behalf. I have personal knowledge of the facts set forth herein, and if called upon to do so I could testify competently to them.

2. My professional qualifications are set forth in the Declaration previously filed in support of Plaintiff's Motion for Class Certification.

3. In addition to the documents previously reviewed, I have since reviewed Intellicorp's Memorandum in Opposition to Plaintiff's Motion for Class Certification and the accompanying Declarations of Ms. Sebeck and Mr. Garrett, as well as Defendant's Responses to Plaintiff's Second Set of Special Interrogatories.

Mr. Garrett Overstates The Size of the Dataset To Be Processed

4. In his declaration in support of Intellicorp's opposition to Plaintiff's motion for class certification, Mr. Garrett states that Intellicorp "has fulfilled more than five million" requests for criminal background information from its instant criminal databases over the period April 16, 2007 to October 18, 2012, which approximates the class period in this case. (Garrett Decl. ¶ 2). However, for the purpose of ascertaining individuals in the class proposed in this case, this number is both irrelevant and misleading.

5. As defined, the proposed class only includes those instances where both a Criminal SuperSearch and Single County Criminal search were run. Based on Intellicorp's Responses to Plaintiff's Special Interrogatories Numbers 14 and 26, Intellicorp already demonstrated its ability to identify (and has identified) the maximum number of reports in the proposed class through October 18, 2012. According to Response Number 26, Intellicorp furnished a total of 838,721 Criminal SuperSearch results over that period. And, according to Response Number 14, Intellicorp furnished 586,440 Criminal SuperSearch results during the same period where its clients did not also request a corresponding Single County Criminal search. Therefore, the maximum number of reports at issue for that period does not exceed 252,281. Mr. Garrett's calculations of the time it would take to process the records is, in contrast, based on 1.5 million records (Garrett Decl. ¶ 30), and is thus greatly overstated.

6. In my previous declaration in support of Plaintiff's motion for class certification, I did not address the wide variety of inconsistencies in the data claimed by Mr. Garrett in his Declaration, and based my estimates on an assumption that such

inconsistencies are not widespread in the dataset. However, I am generally familiar with the problem of parsing inconsistently-formatted data and routinely craft ways to solve it in my day-to-day work. It is a real-world problem, so solutions must also be real-world.

7. In short, what effectively can be parsed by a script is parsed by a script. What cannot be parsed by a script is then left for humans to manually parse. The issue is not whether the entire dataset can be parsed by a script (using fuzzy logic or string parsing functions), but rather whether a script can effectively reduce the number of data to be parsed manually. Because I do not have the actual dataset to review, I cannot provide a detailed solution to each of the problems asserted by Mr. Garrett in his declaration. However, based on my past experiences, I am highly confident that the number of records that might require manual parsing can be substantially reduced and that the problem can be solved with relatively little investment of time and resources.

8. Mr. Garrett states that there are more than 500 sources that feed data into Intellicorp's instant criminal databases, but that is not a relevant statistic. In order to load the data it receives from the various sources into Intellicorp's database, Intellicorp's "data loaders" use ETL (Extract-Transfer-Load) programs to "map" the data to the specified tables and fields. The inquiry at hand focuses on the data *after* it has been loaded into those destinations. Further, and as discussed below, only a subset of the sources would present the problem of having to parse inconsistently-formatted text data. Therefore, the number of records that would require more detailed parsing is substantially smaller than 252,281.

Mr. Garrett Overstates The Complexity Of Parsing Inconsistently-Formatted Data

9. While Mr. Garrett claims that bulk data populate over 150 fields in Intellicorp's instant FCRA criminal database, only a small number of those fields are relevant in this case. (Garrett Decl. ¶ 6).

10. First, the approximately 252,281 records of the order results table at issue can easily be identified through Intellicorp's existing relational database using the ProductId and Order tables and then loaded into a separate database for our analysis. Once that is done only a section of each xml file (stored in the ResultsXML field), needs to be shredded into individual tables. The analysis is then primarily done based on those individual tables. These tables will consist of about 26 fields, of which only 6 fields (Cases.Sentence, Cases.CaseStatus, Charges.ChargeCode, Charges.ChargeDesc, Charges.Disposition, and Charges.OffenseLevel) have a DataType of variable text with limited length, and only 2 (Cases.CaseNotes and Charges.ChargeSentence) have a text DataType of unlimitted text. The 6 fields with variable text data type require some simple parsing, while the 2 text fields may require some additional parsing. This is significantly fewer than the 150 fields Mr. Garrett claims would require customized parsing. Not all 8 fields will have data at all times thus only a subset of the data would need to be parsed.

11. Intellicorp's FCRA databases have 53 disposition tables. When Mr. Garrett refers to over 38,000 distinct disposition records, he sums all disposition records across all 53 FCRA State databases. However, the number of distinct values within any single State does not equal the total sum of 38,299; it is a fraction of that. The matter of fact is that over 80% of the individual states have less than 200 distinct disposition codes. To determine if any one consumer's Criminal SuperSearch results match their Single County Criminal results, we need only to compare the disposition codes for the specific case in question.

12. Matching records is foundational to all SQL servers. A common requirement is to match data by identifying records that are similar but not necessarily exactly the same (due to spelling mistakes for example). In order to do this efficiently we can use the built in scripts and write new ones. Here is one of the feasible ways to do this efficiently.

A. Eliminate quickly a majority of records by utilizing the database server's built in function, such as fuzzy matching.

B. Identify patterns for which to write custom routines for the set of records that did not pass the threshold set for the fuzzy matching (details explained below).

C. Manually compare and match (or not match) the remaining data group.

13. The Database Server has a built in function called fuzzy lookup. By using this and if needed another string Metric Function we can get a high degree of accuracy in matching different but similar values. The number of matches can be adjusted by degree of similarity and confidence. See the example below that are the results of an example that utilized Fuzzy Logic with specific thresholds for Similarity and Confidence

Records that passed the adjustable threshold (matched):

ID	PartName	_Similarity_Name	_Similarity	_Confidence
1	Lock Nut 5	Lock Nut 5	1	1
2	HL Road Seat/Saddle	HL Road Seat/Saddle	1	1
3	HL Mountain Seat Assembly	HL Mountain Seat Assembly	1	1
4	Thin-Jam Hex Nut 1	Thin-Jam Hex Nut 1	1	1
5	Lock Nut 12	Lock Nut 12	1	1
6	HLRoad Rear Wheel	HL Road Rear Wheel	0.9503208	0.9875
7	Hex Nut 7	Hex Nut 7	1	1
8	Thin-Jam Lock Nut 6	Thin-Jam Lock Nut 6	1	1
9	Lock Ring	Lock Ring	1	1
10	Metal Treadl Plate	Metal Tread Plate	0.9462852	0.9875
11	Internal Lock Washer 3	Internal Lock Washer 3	1	1
12	Lock Nut #19	Lock Nut 19	0.9875	0.8647837

Records that did not pass the chosen and adjustable thresholds:

ID	PartName	_Similarity_Name	_Similarity	_Confidence	_Match
1	HLGrip Tpe	HL Grip Tape	0.7986743	0.7989103	LIKELY
2	Lock washar 5	Lock Washer 5	0.93258	0.6788168	NON-MATCH
3	Thin-Jam Lock Nut #2	Thin-Jam Lock Nut 2	0.9875	0.734405	NON-MATCH
4	Seat-Tube	Seat Tube	0.9875	0.5	NON-MATCH
5	LL Moutnain Assembly	LL Mountain Assembly	0.923125	0.6139358	NON-MATCH

Records (values) that have been matched can be "set aside" and simple custom scripts can be written for remaining records that have repeating patterns. A very small percentage of the records that remain after this could be manually matched or confirmed not to be matched.

14. As for inconsistencies in case numbers, Mr. Garrett overstates the problem by implying that all 500 sources for Criminal SuperSearch and all 3,000 sources for Single County Criminal searches must be taken into account when comparing case numbers. (Garret Decl. ¶ 18). However, this is not the case. The issue, rather, is comparing case numbers reported on one Criminal SuperSearch result with case numbers reported on an associated Single County Criminal search result. In this case, any number of basic techniques (such as stripping all non-alphanumeric characters and then identifying the longest common substring) would allow a programmer to match the case numbers with a high degree of accuracy and a great degree of confidence, even if they are formatted differently

15. As for inconsistencies in county names and charge levels, I believe that it is quite possible to match them despite the claimed inconsistencies, based on my previous experiences with various datasets as discussed in more detail below.

My Experiences With Parsing A Large Volume of Inconsistently-Formatted Data

16. I am familiar with the practice of "dumping" unparsed data into text fields and parsing such dumped data on a large scale. The process of parsing generally proceeds in the same manner, where I analyze the entire dataset to identify patterns, write scripts to parse what effectively can be parsed, and identify what cannot easily be parsed. The remainder of the data that cannot easily be parsed is reserved for manual parsing.

17. For example, one of the projects I have recently completed was to parse data containing address and non-address information into proper address fields such as street, city, state, zip, etc. A wide variety of inconsistencies existed in over 10 million records. However, using the technique described above, I was able to correctly parse all

but 5,000 records (or 0.05% of all records) with a set of scripts written over a period of 1.5 days. These 5,000 records too were correctly parsed with a set of additional scripts so that only a couple hundred records needed to be processed manually. The entire process took one developer (*i.e.* me) just 2 days. While I have not viewed Intellicorp's dataset and therefore cannot be certain, a reasonable assumption of 0.1% of data not easily parsed by a script results in 250 records that would have to be manually parsed.

18. Another example is when I had to parse policy numbers received from numerous insurance companies, each with its own proprietary format and "spelling" errors. This number was stored in parsed sections (prefix, suffix, body, certification number, inception date, etc.) and had to be matched against a policy string identifier. There were a total of 15 million records, which required a writing of 50 "scripts" or subroutines. It took me no more than 2 days to write these 50 subroutines.

19. Mr. Garrett explicitly admits that Intellicorp is able to "regenerate, through its transactional database, the full and final reports (assuming all searches were completed at the time of regeneration) containing the results of all searches that were ordered by the customer." (Garrett Decl. ¶¶ 14, 23). I believe that "the results of all searches that were ordered by the customer" as referenced by Mr. Garrett are the XML data I referenced in my previous declaration as stored in ResultXML column in OrderResults table in MasterSearch database on Hercules.

20. Mr. Garrett states that I was incorrect when I previously stated that "On the actual report itself, the value of "NOT PROVIDED" is displayed for a field when the search returns no value for the field." He then states that "This is only correct with respect to certain fields of information." (Garrett Decl. ¶ 25). I believe that my previous statement holds true for all *relevant* fields, *i.e.* the "core fields" as identified by Intellicorp, such as "disposition" field.

21. Mr. Garrett estimates that "it would take between 1 and 2 weeks of dedicated processing time to process [1.5 million] records." (Garrett Decl. ¶ 30). In

addition to the fact that I do not agree that 1.5 million records would have to be processed, I do not agree with that estimate.

22. Last September my team loaded data from 20 different flat files into staging tables with a total record count of 93,832,159 records in 1 hour and 12 minutes. This is about 1 million records per minute.

23. Last August I loaded 53 flat files (20GB of Data) into 99 tables in 1 hour and 24 minutes that resulted in the creation of 329,138,172 records. This is almost 4 million records per minute.

24. Currently I am loading on a daily basis about 1.7 million records from flat files, into designated tables while the data is scrubbed, validated and processed in about 5 minutes. Even with these extra processing requirements, I am loading about 300,000 records per minute.

25. It is difficult to compare the number of records migrated or copied among diverse systems because of the dependencies of hardware (SAN/HardDrives, Memory, CPU), software and the optimal use of them. However, partially shredding 250.000 xml records (stored in a table) into 3 relational data tables should, in my professional opinion, take less than an hour to complete. In any case, Mr. Garrett's prediction is based on processing 1.5 million records. Even using this estimate, processing 250,000 records would take only 1/6 the time, i.*e.* between 1 and 3 days.

I declare under the penalty of perjury under the laws of the United States of America that the above is true and correct. Executed March 15, 2013 in Napa, California.

Henk Valk